How to pretend that correlated variables are independent by using difference observations

Citation for published version:
Williams, CKI 2005, 'How to pretend that correlated variables are independent by using difference observations' Neural Computation, vol. 17, no. 1, pp. 1-6. DOI: 10.1162/0899766052530884

Digital Object Identifier (DOI):
10.1162/0899766052530884

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Neural Computation

NOTE Communicated by Heiga Zen

How to Pretend That Correlated Variables Are Independent by Using Difference Observations

Christopher K. I. Williams
c.k.i.williams@ed.ac.uk
School of Informatics, University of Edinburgh, Edinburgh EH1 2QL, U.K.

In many areas of data modeling, observations at different locations (e.g., time frames or pixel locations) are augmented by differences of nearby observations (e.g., δ features in speech recognition, Gabor jets in image analysis). These augmented observations are then often modeled as being independent. How can this make sense? We provide two interpretations, showing (1) that the likelihood of data generated from an autoregressive process can be computed in terms of "independent" augmented observations and (2) that the augmented observations can be given a coherent treatment in terms of the products of experts model (Hinton, 1999).

1 Introduction

In automatic speech recognition, it is often the case that hidden Markov models (HMMs) are used on observation vectors that are augmented by difference observations (so-called δ features; see Furui, 1986). Under the HMM, each observation vector is modeled as being conditionally independent given the hidden state. How can this make sense, as close-by differences are clearly not independent? A similar difficulty arises in image analysis tasks such as texture segmentation (see, e.g., Dunn & Higgins, 1995). Here derivative features obtained from Gabor filters or wavelet analysis, for example, are modeled as being independent at different locations, despite the fact that these features will have been computed sharing some pixels in common.

In this article, we present two solutions to this problem. In section 2, we show that if the data are generated from a vector autoregressive (AR) model, then the likelihood can be expressed in terms of "independent" difference observations. In section 3, we show that the local models at each location can be combined using a product of experts model (Hinton, 1999) to provide a well-defined joint model for the data and that this can be related to AR models. Section 4 discusses how these interpretations are affected if the local models are conditional on a hidden state variable, as is the case for HMMs.

2 An AR Model

Consider a temporal vector autoregressive model,

$$\mathbf{X}_t = \sum_{i=1}^{p} A_i \mathbf{X}_{t-i} + \mathbf{N}_t, \tag{2.1}$$

where the A_i's are square matrices and $\mathbf{N}_t$ is independent and identically distributed gaussian noise $\sim N(\mathbf{0}, \Sigma_{\mathbf{N}})$. $\mathbf{X}_t$ and $\mathbf{N}_t$ have dimension D for all t. To avoid complicated end effects, we will use periodic (wraparound) boundary conditions, so that the subscript $t-i$ should be read $\mathrm{mod}(t-i, N)$. Thus, there are N random variables $\mathbf{X}_0, \ldots, \mathbf{X}_{N-1}$, which collectively we denote as $\mathbf{X}$, and similarly for $\mathbf{N}$. Then $\mathbf{X}$ and $\mathbf{N}$ are related by $\mathbf{N} = T\mathbf{X}$ for an appropriate matrix T. Thus,

$$P(\mathbf{X}) \propto \prod_{t=0}^{N-1} \exp\left\{-\frac{1}{2}\mathbf{N}_t^T \Sigma_{\mathbf{N}}^{-1} \mathbf{N}_t\right\} \tag{2.2}$$

$$= \prod_{t=0}^{N-1} \exp\left\{-\frac{1}{2}\left[\sum_{i=0}^{p} A_i \mathbf{X}_{t-i}\right]^T \Sigma_{\mathbf{N}}^{-1} \left[\sum_{i=0}^{p} A_i \mathbf{X}_{t-i}\right]\right\}, \tag{2.3}$$

where we have set $A_0 = -I$ so that $\mathbf{N}_t = -\sum_{i=0}^{p} A_i \mathbf{X}_{t-i}$.

Now let $\mathbf{Y}_t^0, \ldots, \mathbf{Y}_t^p$ be linearly independent linear combinations of $\mathbf{X}_t, \ldots, \mathbf{X}_{t-p}$. For example, we could choose $\mathbf{Y}_t^0 = \mathbf{X}_t$, $\mathbf{Y}_t^1 = \mathbf{X}_t - \mathbf{X}_{t-1}$ and so on. As the $\mathbf{Y}_t^i$'s are simple linear combinations of $\mathbf{X}_t, \ldots, \mathbf{X}_{t-p}$, we have

$$\sum_{i=0}^{p} A_i \mathbf{X}_{t-i} = \sum_{i=0}^{p} B_i \mathbf{Y}_t^i, \tag{2.4}$$

for some set of matrices B_i. We can now write

$$P(\mathbf{X}) \propto \prod_{t=0}^{N-1} \exp\left\{-\frac{1}{2}\left[\sum_{i=0}^{p} B_i \mathbf{Y}_t^i\right]^T \Sigma_{\mathbf{N}}^{-1} \left[\sum_{i=0}^{p} B_i \mathbf{Y}_t^i\right]\right\}, \tag{2.5}$$

showing that the likelihood of the underlying $\mathbf{X}$ process can be expressed in terms of a product of terms involving the difference observations up to order p at each time. Stacking $\mathbf{Y}_t^0, \mathbf{Y}_t^1, \ldots, \mathbf{Y}_t^p$ as the vector $\mathbf{Y}_t$ we have

$$P(\mathbf{X}) \propto \prod_{t=0}^{N-1} \exp\left\{-\frac{1}{2}\mathbf{Y}_t^T M \mathbf{Y}_t\right\}, \tag{2.6}$$

where the (i, j) block of the matrix M (between $\mathbf{Y}_t^i$ and $\mathbf{Y}_t^j$) has the form $B_i^T \Sigma_{\mathbf{N}}^{-1} B_j$. Equation 2.6 almost looks like a product of independent gaussians, but note that M is singular (it has rank D as it arises from $\mathbf{N}_t$) so the correct normalization factor of the gaussian cannot be obtained from it.

As a simple example, consider the scalar AR(1) process $X_t = \alpha X_{t-1} + N_t$ and set $Y_t^0 = X_t$, $Y_t^1 = X_t - X_{t-1}$. Thus,

$$X_t - \alpha X_{t-1} = (1-\alpha)X_t + \alpha(X_t - X_{t-1}) \tag{2.7}$$

$$= (1-\alpha)Y_t^0 + \alpha Y_t^1. \tag{2.8}$$

To obtain the likelihood for the sequence X, the matrix M will have the form

$$M = \frac{1}{\sigma_n^2}\begin{pmatrix} (1-\alpha)^2 & \alpha(1-\alpha) \\ \alpha(1-\alpha) & \alpha^2 \end{pmatrix}, \tag{2.9}$$

where $\sigma_n^2 = \mathrm{var}(N_t)$. As expected, M has rank 1 (it is an outer product).

Interestingly, the matrix M is not equal to the inverse covariance of the $\mathbf{Y}_t$'s derived from the distribution for $\mathbf{X}$. To show this, we first use the result that for the scalar AR(1) process on the circle, the covariance $C[j] = \langle X_t X_{t-j}\rangle$ is given by

$$C[j] = \frac{\sigma_n^2(\alpha^{|j|} + \alpha^{|N-j|})}{(1-\alpha^2)(1-\alpha^N)}. \tag{2.10}$$

Thus,

$$\mathrm{cov}(\mathbf{Y}_t) = \begin{pmatrix} \langle Y_t^0 Y_t^0\rangle & \langle Y_t^0 Y_t^1\rangle \\ \langle Y_t^0 Y_t^1\rangle & \langle Y_t^1 Y_t^1\rangle \end{pmatrix} = \begin{pmatrix} C[0] & (C[0]-C[1]) \\ (C[0]-C[1]) & 2(C[0]-C[1]) \end{pmatrix}. \tag{2.11}$$

Inversion of $\mathrm{cov}(\mathbf{Y}_t)$ shows that it is not equal to M as given in equation 2.9. Notice that the joint distribution of $\mathbf{Y}_0, \ldots, \mathbf{Y}_{N-1}$ is singular.

If we take an AR process on the $\mathbf{X}$ variables, then one can choose linear combinations of the $\mathbf{X}_t$s that are truly independent by carrying out an eigenanalysis. (For the periodic boundary conditions described above and time-invariant coefficients, the eigenbasis would be the Fourier basis.) However, if we allow ourselves an overcomplete basis set, then we have shown that the likelihood of $\mathbf{X}$ under the AR process can readily be computed using "independent" densities at each location.

Although we have given the derivation above using gaussian noise, in fact the conclusion concerning expressing the likelihood of the $\mathbf{X}$ sequence in terms of a product of terms involving $\mathbf{Y}_t$'s is independent of the form of the noise driving the AR process.

It is also possible to extend the AR model described above beyond the temporal one-dimensional chain. For example, Abend, Harley, and Kanal (1965) describe Markov mesh models in two dimensions. A simple example of such a model is a "third-order" Markov mesh, where $\mathbf{X}_{i,j}$ depends autoregressively on $\mathbf{X}_{i,j-1}$, $\mathbf{X}_{i-1,j-1}$, and $\mathbf{X}_{i-1,j}$. The same construction in terms of $\mathbf{Y}$ variables can be used in this case.

3 Product of Experts Interpretation

At an individual location, we have a model $P_t(\mathbf{Y}_t)$ for the augmented vector $\mathbf{Y}_t$. To define a joint distribution on $\mathbf{X}$, we set

$$P(\mathbf{X}) = \frac{1}{Z}\prod_t P_t(\mathbf{Y}_t), \tag{3.1}$$

where Z is a normalization constant (known in statistical physics as the partition function). This is the product of experts construction (Hinton, 1999). One can also think of this as a Markov random field construction where $P(\mathbf{X}) \propto \exp -E(\mathbf{X})$ and $E(\mathbf{X}) = -\sum_t \log P_t(\mathbf{Y}_t)$. If each $P_t(\mathbf{Y}_t)$ is gaussian, then $P(\mathbf{X})$ will also be gaussian, and $Z = (2\pi)^{N/2}|C|^{1/2}$ where C is the covariance matrix of $\mathbf{X}$.

Again we consider a simple example relating to a scalar AR(1) process, so $\mathbf{Y}_t = (X_t, X_t - X_{t-1})^T$. Let

$$P_t(\mathbf{Y}_t) \propto \exp -\frac{1}{2}\{a_0 X_t^2 + a_1(X_t - X_{t-1})^2\}, \tag{3.2}$$

with $a_0, a_1 > 0$. Then we obtain the joint distribution,

$$P(\mathbf{X}) \propto -\frac{1}{2}\left\{a_0\sum_t X_t^2 + a_1\sum_t (X_t - X_{t-1})^2\right\}. \tag{3.3}$$

C^{-1}, the inverse covariance matrix of $\mathbf{X}$, is circulant with entries $a_0 + 2a_1$ on the diagonal and $-a_1$ in the bands above and below the diagonal and in the northeast and southwest corners. For the AR(1) process, $X_t = \alpha X_{t-1} + N_t$ with $N_t \sim N(0, \beta^{-1})$, we obtain corresponding entries of $\beta(1+\alpha^2)$ on the diagonal and $-\beta\alpha$ off the diagonal. The overall scale of a_0 and a_1 has the same effect as β in setting the variance of the process but $r \overset{def}{=} \frac{a_0}{a_1} = \frac{(1-\alpha)^2}{\alpha}$, so for any given α value, there is a corresponding value of r[1].

For the gaussian case with expert t involving interactions between $\mathbf{X}_t$ and $\mathbf{X}_{t-p}$, we obtain a quadratic form with the same pattern of banding as in the inverse covariance matrix of an AR(p) process, but as above, for some choices of parameters there may not be a corresponding AR process.

Again this construction can be extended to two (or more) dimensions. For example, in 2D we might consider the variable $\mathbf{X}_{i,j}$ and the differences to its four neighbors to the north, south, east, and west to obtain a five-dimensional $\mathbf{Y}$ vector. Equation 3.1, with each expert being gaussian, then defines a gaussian Markov random field over the lattice of $\mathbf{X}$ variables.

[1] Interestingly for $r \in (-4, 0)$, there are no corresponding values of α. Note that $\alpha = 0 \Rightarrow a_1 = 0$.

4 Incorporating Hidden State

In speech recognition using HMMs, the $\mathbf{Y}_t$s are modeled as conditionally independent given the discrete hidden variable s_t. We now consider how this affects the interpretations given above.

For interpretation 1, we consider a switching AR(p) process or AR-HMM (see, e.g., Woodland, 1992), so that $\mathbf{X}_t$ depends on $\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-p}$ and also s_t. For example, using gaussian noise and setting $s_t = k$, we have $\mathbf{X}_t \sim N(\sum_{i=1}^{p} A_i^k \mathbf{X}_{t-i}, \Sigma^k)$. Notice that the AR model parameters now depend on the switching variable. However, we can still write the prediction $\sum_{i=1}^{p} A_i^k \mathbf{X}_{t-i}$ as a linear combination of the $\mathbf{Y}_t^i$s, so the likelihood can be written in the form of "independent" contributions from the $\mathbf{Y}_t$s. Note that the usual forward and backward HMM recursions can be carried out for the AR-HMM.

For interpretation 2, we have the individual component densities $P_t(\mathbf{Y}_t \mid s_t)$, and the joint distribution

$$P(\mathbf{X} \mid \mathbf{s}) = \frac{1}{Z(\mathbf{s})} \prod_t P_t(\mathbf{Y}_t \mid s_t), \tag{4.1}$$

where $\mathbf{s} = (s_0, \ldots, s_{N-1})$. Notice that the normalization constant in general depends on $\mathbf{s}$, and thus when given $\mathbf{X}$, the computation of $P(\mathbf{X} \mid \mathbf{s})$ depends no only on the component densities but also on $Z(\mathbf{s})$. However, if $P_t(\mathbf{Y}_t \mid s_t)$ is gaussian and has the same covariance structure but different means depending on s_t for all t, then Z would turn out to be independent of $\mathbf{s}$.

While writing this article, I became aware of the work of Tokuda, Zen, and Kitamura (2003), who correctly derive the product of gaussian experts construction conditional on $\mathbf{s}$ and note the general dependence of $Z(\mathbf{s})$ on $\mathbf{s}$. They also observe that use of the Viterbi algorithm to find the state sequence $\mathbf{s}$ that maximizes $P(\mathbf{s}) \prod_t P_t(\mathbf{Y}_t \mid s_t)$ (which is easily done with standard dynamic programming techniques) will not, in general, yield the sequence that maximizes $P(\mathbf{s} \mid \mathbf{X})$, because of the $Z(\mathbf{s})$ term.

Most practical HMM-based speech recognition systems use mixtures of gaussians to model the $\mathbf{Y}_t$s at each frame. The product of experts interpretation readily handles this situation. For an AR model interpretation, the use of a mixture distribution for the $\mathbf{Y}_t$s already suggests a switching AR process with the switching variable hidden.

5 Discussion

We have described both conditionally specified models (AR processes) and simultaneously specified models (products of experts) to define the joint density[2] $P(\mathbf{X})$ and relate it to the augmented feature vectors $\{\mathbf{Y}_t\}$.

[2] This terminology is derived from Cressie (1993, sec. 6.3).

While this article describes a theoretical framework for understanding why using difference observations make sense, it would be interesting to examine empirically the question of how well AR and products of experts models characterize the dependencies between time frames or pixel locations.

Acknowledgments

This note was inspired by questions raised by Joe Frankel's Ph.D. thesis. Thanks to John Bridle and Joe Frankel for helpful conversations and comments on earlier drafts, to Joe Frankel for drawing my attention to Tokuda et al. (2003), and to the anonymous referees for their comments, which helped to improve the article.

References

Abend, K., Harley, T. J., & Kanal, L. N. (1965). Classification of binary random patterns. *IEEE Transactions on Information Theory, 11*(4), 538–544.

Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: Wiley.

Dunn, D., & Higgins, W. E. (1995). Optimal Gabor filters for texture segmentation. *IEEE Transactions on Image Processing, 4*(7), 947–964.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing, 34*, 52–59.

Hinton, G. E. (1999). Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks* (Vol. 1, pp. 1–6). London: IEE.

Tokuda, K., Zen, H., & Kitamura, T. (2003). Trajectory modelling based on HMMs with the explicit relationship between static and dynamic features. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*. N.P.: International Speech Communication Association.

Woodland, P. C. (1992). Hidden Markov models using vector linear prediction and discriminative output distributions. In *Proceedings of 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. I, pp. 509–512). Piscataway, NJ: IEEE.

Received February 3, 2004; accepted May 28, 2004.